MANTA: Physics-Informed Generalized Underwater Object Tracking

Suhas Srinath[1] *   Hemang Jamadagni[2] *   Aditya Chandrasekar[1] †   Prathosh A P[1]

[1]Indian Institute of Science   [2]National Institute of Technology Karnataka

Abstract

Underwater object tracking is challenging due to wavelength-dependent attenuation and scattering, which severely distort appearance across depths and water conditions. Existing trackers trained on terrestrial data fail to generalize to these physics-driven degradations. We present MANTA, a physics-informed framework integrating representation learning with tracking design for underwater scenarios. We propose a dual-positive contrastive learning strategy coupling temporal consistency with Beer–Lambert augmentations to yield features robust to both temporal and underwater distortions. We further introduce a multi-stage pipeline augmenting motion-based tracking with a physics-informed secondary association algorithm that integrates geometric consistency and appearance similarity for re-identification under occlusion and drift. To complement standard IoU metrics, we propose Center–Scale Consistency (CSC) and Geometric Alignment Score (GAS) to assess geometric fidelity. Experiments on four underwater benchmarks (WebUOT-1M, UOT32, UTB180, UWCOT220) show that MANTA achieves state-of-the-art performance, improving Success AUC by up to 6%, while ensuring stable long-term generalized underwater tracking and efficient runtime. Code available at https://github.com/Kazedaa/MANTA.

1. Introduction

Underwater object tracking (UOT) poses challenges distinct from terrestrial tracking due to severe image degradations from wavelength-dependent attenuation, scattering, and color absorption [12, 18, 19, 50]. These factors drastically alter object appearance across depths and conditions, creating a significant domain gap that hinders a direct transfer from terrestrial trackers [20, 34, 61]. Similar to general tracking, UOT includes single-object tracking (SOT) and multi-object tracking (MOT), but progress has been driven largely by SOT [2, 3, 42, 56, 57] given the scarcity of MOT datasets [43]. Most existing UOT approaches adapt terrestrial methods via preprocessing or fine-
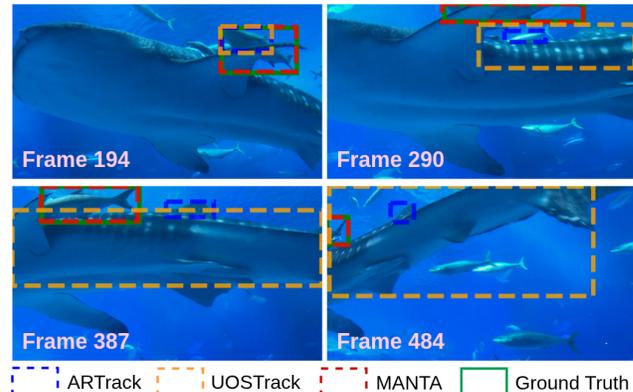
---

*Equal contribution. †Work done while at Indian Institute of Science.

Figure 1. Tracking outputs from different methods on `Video_0002` from the UTB180 [2] dataset. MANTA (dashed red) reliably tracks the target correct object (solid green) across frames, maintaining identity even under occlusion or temporary disappearance.

tuning [3, 33, 56], yet such methods fail to capture physics-driven distortions inherent to underwater imaging. As a result, trackers struggle with wavelength-selective absorption, scattering, and depth-dependent contrast loss. This work focuses on advancing underwater SOT by explicitly modeling these domain-specific challenges.

Recent advances in self-supervised learning (SSL) offer a promising alternative, enabling the learning of powerful visual representations without reliance on labeled data [25]. In particular, contrastive learning has demonstrated strong capability in learning semantically meaningful and generalizable features by enforcing consistency across differently augmented views of the same instance [11, 31, 44]. Yet, most SSL pipelines employ domain-agnostic augmentations such as random cropping, color jittering, or geometric transformations. These transformations fail to capture underwater-specific distortions, limiting their applicability to UOT.

To address these challenges, we introduce a physics-informed self-supervised framework for underwater SOT. Our dual-positive contrastive learning strategy combines temporal consistency with physics-based priors by constructing positive pairs from (1) temporally aligned object

crops across frames and (2) Beer–Lambert [49] augmented crops simulating depth-dependent attenuation. This enforces invariance to both temporal and underwater distortions, yielding domain-generalizable features for UOT.

Beyond representation learning, robust tracking must address occlusion, disappearance, and drift. Lightweight real-time trackers perform well terrestrially [9, 51, 58] but degrade underwater due to overlapping objects, domain gaps, and frequent identity loss [38]. We mitigate these issues by augmenting a motion-based primary tracker with a physics-informed secondary association algorithm that leverages visual cues for long-term re-identification, improving temporal coherence and stability under challenging conditions.

Finally, we introduce two metrics for finer-grained evaluation: Center–Scale Consistency (CSC), evaluating both positional accuracy and bounding-box scale, and Geometric Alignment Score (GAS), applying Gaussian penalties to jointly capture center and scale misalignment. In summary, we present **MANTA** (**MA**ri**N**e object **T**racking **A**lgorithm), a physics-informed self-supervised framework for underwater SOT that integrates robust representation learning into a physics-informed vision-based secondary association algorithm for efficient and generalizable tracking. Our contributions are summarized as follows:

- We propose a dual-positive contrastive learning framework that combines temporal consistency with Beer–Lambert augmentations, enabling features independent of both temporal variations and underwater distortions.
- We design a multi-stage tracker that augments motion cues with a physics-informed secondary association using geometric consistency and appearance similarity. Cascaded strategies of ID reuse, geometric scoring, and localized appearance search ensure stable long-term trajectories under occlusion, clutter, and low visibility.
- We propose two new evaluation metrics: Center-Scale Consistency (CSC), requiring accurate center and scale prediction, and Geometric Alignment Score (GAS), applying Gaussian penalties for misalignment to provide finer tracker evaluation insights beyond IoU.
- Extensive experiments on four underwater benchmarks show consistent state-of-the-art performance, with improvements in Success AUC, geometric alignment, and long-term stability.

2. Related Work

2.1. Underwater Object Tracking

Underwater object tracking is critical for marine robotics, autonomous vehicles, and biology [4, 20, 32, 39]. Early adaptations of terrestrial trackers using preprocessing such as histogram equalization and color correction [26]), struggle with underwater distortions. Later designs introduced underwater-specific correlation filters [50] and adaptive fea-

ture selection [32], but these hand-crafted methods fail to generalize across environments. Deep learning–based trackers [6, 8, 14, 15, 30] improved performance, yet their reliance on terrestrial pretraining (e.g., ImageNet [17], COCO [36], GOT-10k [23]) created domain gaps. Classical real-time trackers [9, 51, 58] are efficient but often fail under occlusion or when re-identifying lost objects. Our work addresses these gaps by learning domain-robust features via physics-informed self-supervised learning, coupled with a secondary vision-guided association algorithm that complements motion-based tracking for improved re-identification.

2.2. Self-Supervised Learning for Visual Tracking

Self-supervised learning has enabled robust representation learning without manual annotations [11, 31, 44], and has recently been explored for visual tracking [22]. Contrastive learning, in particular, has shown strong potential by constructing positive pairs from temporal sequences or spatial augmentations to enforce independence [55, 60]. For instance, De Plaen et al. [16] showed that contrastive training on tracking sequences can outperform ImageNet pretraining, while other works leverage object crops across frames [13] or optical flow correspondences [35]. However, most self-supervised tracking methods rely solely on temporal consistency and generic augmentations, overlooking domain-specific information. In contrast, our approach integrates temporal contrastive learning with physics-informed augmentations (depth-based), yielding a dual-supervision framework that captures both global object identity and is robust to underwater distortions.

2.3. Physics-Informed Underwater Vision

The integration of physical principles into machine learning has gained significant attention across scientific domains [7, 27]. In computer vision, physics-based models have been applied to image restoration [46], depth estimation, and domain adaptation [59]. Central to underwater imaging is the Beer–Lambert law [40, 49], which models wavelength-dependent light attenuation and provides a principled description of image degradation caused by absorption, scattering, and backscattering [1]. Prior works have primarily exploited this physics for enhancement and restoration [10, 48], including Dark Channel Prior adaptations [12] and GAN-based underwater-to-terrestrial mappings [5]. In contrast, our work employs these physical models not for image correction, but as augmentations within a self-supervised framework. By embedding Beer–Lambert–based transformations into temporal contrastive learning, we explicitly learn features independent of underwater distortions while preserving semantic object identity. This strategy is different from generic augmentation pipelines and enables the learning of domain-independent features that generalize effectively to underwa-
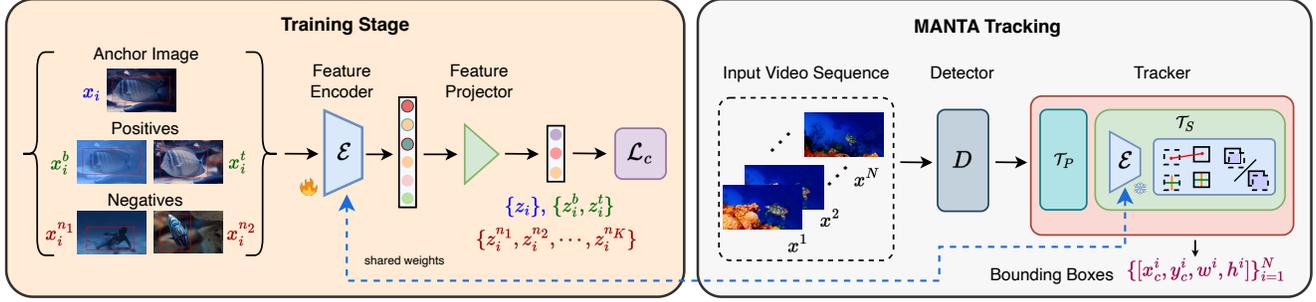
Figure 2. Overview of MANTA. The self-supervised encoder $\mathcal{E}$ is trained via contrastive learning with Beer–Lambert augmentations x_i^b and temporal augmentations x_i^t. Following detections from D and primary tracking with $\mathcal{T}_P$, embeddings produced by $\mathcal{E}$ are used for vision-guided secondary association $\mathcal{T}_S$, which refines trajectories by matching IoU, scales, and centers, yielding accurate predictions for an input video sequence.

ter object tracking.

3. Methodology

Our approach addresses the fundamental challenge of underwater object tracking through a physics-informed self-supervised learning framework. The key insight is that underwater imaging follows predictable physical principles, which we can exploit to learn robust representations that generalize across diverse underwater conditions. Our complete pipeline consists of three interconnected components: (1) a dual-positive contrastive learning framework that learns domain-invariant features, (2) physics-based augmentations that simulate realistic underwater distortions, and (3) a multi-stage tracking pipeline that leverages these learned representations for robust object association. Figure 2 provides an overview of our complete methodology.

3.1. Physics-Informed Representation Learning

The core of our approach lies in learning representations that are invariant to physical distortions in underwater environments. We achieve this through a dual-positive contrastive learning strategy that enforces consistency across both temporal variations and physics-based augmentations.

3.1.1. Dual-Positive Contrastive Framework

Traditional contrastive learning methods for tracking [22, 55, 60] primarily exploit temporal correspondence or generic augmentations such as color jitter, motion blur, and saturation. Our framework extends this approach by introducing physics-informed augmentations alongside temporal consistency, creating a dual-supervision mechanism that captures both object identity persistence and robustness to underwater distortions.

For each target object bounding box crop $\mathbf{x}_i$ at time t, we construct two types of positive pairs:

- **Temporal Positives**: Object crops $\mathbf{x}_{t+\Delta t}$ from subsequent frames containing the same object, enforcing temporal consistency of object identity. These crops are denoted as x_i^t with corresponding features z_i^t.
- **Physics Positives**: Beer–Lambert [40, 49] augmented versions of the current crop, simulating depth-dependent underwater distortions. These are denoted as x_i^b with features z_i^b.

Negatives are defined as object crops from other sequences in the batch, denoted by $\{x_i^{n_j}\}_{j=1}^K$ with corresponding features $\{z_i^{n_j}\}_{j=1}^K$. Our contrastive learning objective, based on InfoNCE [41] and supervised contrastive loss [29], encourages similarity between the anchor representation z_i and both temporal and physics positives, while pushing away negatives:

$$\mathcal{L}_c = \frac{1}{N_b} \sum_{i=1}^{N_b} \mathcal{L}_i \qquad (1)$$

where the per-sample loss is defined as

$$\mathcal{L}_i = -\log \frac{\exp(f_\theta(z_i) \cdot f_\theta(z_i^t)/\tau) + \exp\big(f_\theta(z_i) \cdot f_\theta(z_i^b)/\tau\big)}{\sum_{k=1}^K \exp(f_\theta(z_i) \cdot f_\theta(z_i^{n_k})/\tau)}$$

with τ denoting the temperature parameter. This formulation drives the encoder $\mathcal{E}$ to produce features that align representations of the same object across time and varying physical conditions, while maintaining separation from different objects.

3.1.2. Beer-Lambert Physics Augmentations

To generate realistic physics-informed augmentations, we leverage the Beer-Lambert law, which governs light attenuation in underwater environments [40]. This physical model allows us to simulate the depth-dependent appearance variations that objects naturally experience underwater. Given a training image $\mathbf{I}$ and its corresponding depth map $\mathbf{D}$ (obtained using UDepth [54], an efficient off-the-shelf monoc-

ular depth estimator), we compute the Beer-Lambert augmented image as:

$$\mathbf{I}^{BL}(\mathbf{p}) = \mathbf{I}(\mathbf{p}) \cdot T(\mathbf{p}) + \mathbf{B} \cdot (1 - T(\mathbf{p})) \qquad (2)$$

where $\mathbf{p}$ denotes pixel coordinates, $T(\mathbf{p}) = e^{-\beta \mathbf{D}(\mathbf{p})}$ represents the transmission map, β is the attenuation coefficient, and $\mathbf{B}$ represents the background water color. During training, we vary $\beta \in [0.1, 0.5]$ within each batch to simulate diverse depth-dependent attenuation, while keeping a fixed blue–green illumination $\mathbf{B}$. Lower β values yield minimal distortion, whereas higher values risk obscuring objects, ensuring a balance between augmentation diversity and object visibility. By varying β across training batches, the encoder learns to extract features that are invariant to different attenuation levels, enabling generalization across diverse underwater environments without overfitting to any specific water type or depth range. Importantly, Beer–Lambert augmentation serves purely as a training strategy to expose the model to realistic distortions; hence, precise depth accuracy is unnecessary since the objective is invariance across attenuation patterns. Depth maps are used only during training, with inference relying solely on RGB inputs.

3.2. Multi-Stage Tracking Pipeline

Our tracking framework integrates the learned physics-informed representations into a three-stage pipeline that combines the efficiency of motion-based tracking with the robustness of vision-guided association.

3.2.1. Primary Motion-Based Tracking

We begin with OC-SORT [9] as our primary tracker, which provides efficient motion-based association through optical flow consistency and temporal smoothness. While OC-SORT is effective for short-term tracking, it is susceptible to identity switches in cluttered underwater scenes and track loss during occlusions.

3.2.2. Vision-Guided Secondary Association

To address the limitations of motion-only tracking, we introduce a vision-guided secondary association module that leverages our physics-informed features to maintain track continuity. This module operates through a sequential matching strategy designed to recover from track loss while maintaining computational efficiency. Starting from the ground-truth anchor in frame one, the algorithm maintains tracking continuity through the following steps:

1. **Initialization:** In the first frame, we associate the ground-truth anchor with the detection that maximizes a composite geometric score. For two bounding boxes

b_1 and b_2, we define:

$$\text{dist}(b_1, b_2) = 1 - \frac{\|c(b_1) - c(b_2)\|_2}{\max(\text{diag}(b_1), \text{diag}(b_2))} \qquad (3)$$

$$\text{scale}(b_1, b_2) = \frac{\min(\text{area}(b_1), \text{area}(b_2))}{\max(\text{area}(b_1), \text{area}(b_2))} \qquad (4)$$

The composite score combines both spatial and scale consistency as:

$$\text{score}(b_1, b_2) = w_{dist} \cdot \text{dist}(b_1, b_2) + w_{scale} \cdot \text{scale}(b_1, b_2)$$

2. **History-based Track Reuse:** For subsequent frames, we first check if any recently seen track IDs reappear with geometric consistency.
3. **Active Track Retention:** If the current target ID remains active and shows consistency, we retain its bounding box.
4. **Score-based Re-acquisition:** When the target ID is lost, we re-select candidates using the composite score relative to the previous track state.
5. **Physics-Informed Local Search:** As a final fallback, we perform local search around the last known position, using our physics-informed encoder $\mathcal{E}$ to compute cosine similarity between candidate crops and the target representation.

The similarity function for geometric and semantic consistency is defined as:

$$\text{are\_similar}(b_1, b_2) = \mathbb{1}\Big[\text{size\_ok}(b_1, b_2) \wedge \big(\text{IoU}(b_1, b_2) > 0\big)$$
$$\wedge \big(\cos\big(\mathcal{E}(b_1), \mathcal{E}(b_2)\big) \geq \theta_{cos}\big)\Big] \qquad (5)$$

where $\wedge$ denotes logical conjunction (AND), and $\mathcal{E}(b_1)$ and $\mathcal{E}(b_2)$ denote the representations of the object crop with bounding boxes b_1 and b_2, respectively. The $\text{size\_ok}(\cdot, \cdot)$ function and IoU are defined as

$$\text{size\_ok}(b_1, b_2) = \mathbb{1}\left[\frac{\max(w_1, w_2)}{\min(w_1, w_2)} \leq 2 \ \wedge \ \frac{\max(h_1, h_2)}{\min(h_1, h_2)} \leq 2\right]$$

$$\text{IoU}(b_1, b_2) = \frac{|b_1 \cap b_2|}{|b_1 \cup b_2|}$$
$$(6)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, w_1 and w_2 are the widths of the two bounding boxes, and h_1 and h_2 are the heights. The complete algorithm is summarized in Algorithm 1.

3.2.3. End-to-End Integration

Our complete system integrates these components through a three-stage pipeline: (1) **Object Detection** using fine-tuned RF-DETR [45] on curated underwater datasets, (2) **Motion-Guided Association** via OC-SORT for efficient short-term tracking, and (3) **Vision-Guided Re-association** using our

Algorithm 1 Vision-Guided Secondary Association

1: **function** SECONDARY_ASSOCIATION(Primary_Tracks, anchor)
2:     $history.push(-1)$ ▷ LRU Stack
3:     **for** each $frame$ in sequence
4:        $dets = detections[\text{frame}]$
5:        **if** $history.peek(d.id) == -1$
6:           $d^* = \arg\max_{d \in dets} \text{score}(anchor, d.bbox)$
7:           $curr\_bbox = anchor$
8:           $history.push(d^*.id)$
9:        **else**
10:           **for** each $h \in$ history and $d \in dets$
11:              **if** $d.id == h$
12:                 **if** are_similar$(curr\_bbox, d.bbox)$
13:                    $history.push(h)$
14:                    $curr\_bbox = d.bbox$
15:                    **break both loops**
16:           **for** $d \in dets$
17:              **if** $d.id == history.peek(d.id)$
18:                 **if** are_similar$(curr\_bbox, d.bbox)$
19:                    $curr\_bbox = d.bbox$
20:                    **break loop**
21:           $d^* = \arg\max_{d \in dets} \text{score}(anchor, d.bbox)$
22:           $curr\_bbox = d^*.bbox$
23:           $history.push(d^*.id)$
24:           **break loop**
25:           $c^* = \arg\max_{c \in \text{crops}} \cos\left(\mathcal{E}(c), \mathcal{E}(curr\_bbox)\right)$
26:           **if** $\cos\left(\mathcal{E}(c^*), \mathcal{E}(curr\_bbox)\right) \geq \theta_{cos}$
27:              $curr\_bbox \leftarrow c^*.bbox$
28:     append $curr\_bbox$ to predictions
29:     **return** predictions

---

physics-informed features to recover lost targets and maintain long-term consistency. This design leverages the complementary strengths of motion-based efficiency and vision-based robustness, while the physics-informed representations provide the domain-specific invariances necessary for reliable underwater tracking.

4. Experiments and Results

4.1. Experimental Setup and Training Details

All experiments were done on a 48GB NVidia RTX A6000 GPU. The detector was trained using a batch size of $B = 16$, and the physics encoder with a batch size of $N_b = 32$.

Detector Training: We fine-tune RF-DETR [45] on a curated underwater dataset combining SUIM [24], UIIS10k [47], frames from UVOT [3], WebUOT-1M [56] (the official train split), UOT100 [42] (excluding UOT32), and Brackish MOT [43], covering diverse underwater scenes. Missing object class labels were manually anno-

tated, and copy-paste augmentation was used to balance rare classes. The detector was trained for 50 epochs with the standard RF-DETR losses.

Physics Encoder Training: We train the encoder with contrastive learning on frames sampled from WebUOT-1M (train set), UVOT400, and Brackish MOT. Temporal positives x_i^t are generated by sampling frames within four steps of the anchor, while Beer–Lambert augmentations x_i^b use a fixed illumination map $\mathbf{B} = [0.6, 0.8, 0.9]$ (RGB) to model typical blue-green water conditions in SOT datasets. Attenuation strength β is varied in $[0.1, 0.5]$ to simulate depth-dependent distortions. A ResNet-18 [21] backbone with a 128-dimensional projection head is trained using InfoNCE [41] loss and optimized with AdamW [37] at a learning rate of 10^{-5} for 100 epochs.

Tracker-specific parameters: For primary association, we use default OC-SORT [9] parameters. The secondary association uses a sliding window stride factor of 0.5 (half the bounding box dimensions), a search expansion factor of 1.5, $w_{dist} = w_{size} = 1$ and $\theta_{cos} = 0.9$ for the matching of the appearance (lower thresholds led to noisy matches, hindering the re-identification process). These thresholds were selected through validation on a held-out subset of the training data to balance matching precision and recall in the re-identification process.

4.2. Evaluation Datasets

WebUOT-1M [56]: a large-scale benchmark with over 1M frames covering diverse marine organisms, vehicles, and objects under varying water conditions and depths. We report results only on the test set for fair comparison. **UOT32** [28]: 32 challenging sequences featuring fish, divers, and underwater equipment across environments with varying visibility. **UTB180** [2]: 180 high-quality sequences emphasizing difficult cases such as poor visibility, color distortion, and occlusion by marine life. **UW-COT-220** [57]: 220 sequences focusing on camouflaged object tracking, spanning diverse lighting, clarity, and object categories.

4.3. Evaluation Metrics

We evaluate our method using both standard tracking metrics and two novel metrics specifically designed to capture fine-grained tracking performance.

Standard Metrics: We report Success AUC and Precision AUC, which are widely used in visual tracking literature [20]. Success AUC measures the area under the curve of success rates at different overlap thresholds, while Precision AUC measures the area under the curve of precision rates at different center location error thresholds. We also report Success@0.5 and Precision@20px as threshold-specific metrics, along with mean IoU (mIoU).

Center-Scale Consistency (CSC): While conventional precision measures only the alignment of bounding-box

Dataset	Method	Suc. AUC	Suc.@0.5	Pre. AUC	Pre.@20px	mIoU	mGAS	mCSC@0.2
UOT32 [28]	SiamFC	0.4226	0.4823	0.4499	0.4673	0.4955	0.5778	0.1893
	STARK	0.6153	0.7498	0.6432	0.7057	0.6906	0.7671	0.5627
	SiamRPN++	0.6204	0.7636	0.6452	0.6943	0.6782	0.7811	0.5732
	TransT	0.6693	0.8068	0.7097	0.7661	0.7333	0.8239	0.6096
	SeqTrack	0.6581	0.8027	0.7217	0.7932	0.7116	0.8373	0.6075
	OSTrack	0.6640	0.8132	0.7074	0.7685	0.7097	0.8288	0.6195
	OKTrack	0.6307	0.7744	0.6727	0.7448	0.6811	0.7973	0.5852
	ARTrack	0.6612	0.8058	0.7181	0.7769	0.7235	0.8381	0.5984
	UOSTrack	0.6700	0.8300	0.7096	0.7795	0.7285	0.8337	0.6285
	MANTA	0.7306	0.8848	0.7628	0.8419	0.7632	0.8829	0.7438
UTB180 [2]	SiamFC	0.3571	0.3870	0.2470	0.2333	0.5101	0.4836	0.2056
	STARK	0.5439	0.6224	0.4358	0.4547	0.7172	0.6347	0.5066
	SiamRPN++	0.5478	0.6546	0.4243	0.4233	0.6775	0.6856	0.4755
	TransT	0.5852	0.6685	0.4836	0.5086	0.7450	0.6876	0.5458
	SeqTrack	0.6184	0.7099	0.5228	0.5528	0.7830	0.7160	0.6046
	OSTrack	0.6454	0.7380	0.5399	0.5691	0.7611	0.7499	0.6121
	OKTrack	0.7014	0.7999	0.6100	0.6623	0.7860	0.8123	0.7010
	ARTrack	0.7085	0.7980	0.6215	0.6548	0.8021	0.8099	0.6706
	UOSTrack	0.6608	0.7581	0.5588	0.5894	0.7703	0.7689	0.6370
	MANTA	0.7328	0.8268	0.6698	0.7383	0.8351	0.8312	0.7746
UW-COT-220 [57]	SiamFC	0.2830	0.3007	0.1937	0.1743	0.4606	0.3922	0.1580
	STARK	0.4707	0.5345	0.3682	0.3700	0.5968	0.5736	0.3979
	SiamRPN++	0.4268	0.4994	0.3410	0.3363	0.6193	0.5458	0.3313
	TransT	0.4998	0.5729	0.4128	0.4184	0.6686	0.6224	0.4246
	SeqTrack	0.5638	0.6507	0.4934	0.5121	0.6806	0.6849	0.5012
	OSTrack	0.5504	0.6288	0.4677	0.4846	0.6555	0.6638	0.4973
	OKTrack	0.5976	0.6935	0.5162	0.5391	0.6684	0.7205	0.5378
	ARTrack	0.6075	0.6964	0.5385	0.5616	0.6966	0.7283	0.5404
	UOSTrack	0.5747	0.6629	0.4926	0.5116	0.6639	0.6909	0.5204
	MANTA	0.6037	0.7036	0.5255	0.5480	0.7631	0.7376	0.5936
WebUOT-1M [56]	SiamFC	0.3224	0.3487	0.2397	0.2282	0.4728	0.4432	0.1724
	STARK	0.5341	0.6189	0.4507	0.4735	0.6848	0.6467	0.4909
	SiamRPN++	0.4961	0.5928	0.4073	0.4137	0.6255	0.6315	0.4203
	TransT	0.5466	0.6352	0.4754	0.4971	0.6874	0.6721	0.4925
	SeqTrack	0.5647	0.6506	0.5005	0.5244	0.6993	0.6839	0.5239
	OSTrack	0.5919	0.6850	0.5221	0.5477	0.6887	0.7188	0.5491
	OKTrack	0.6294	0.7218	0.5579	0.5943	0.7012	0.7527	0.6031
	ARTrack	0.6393	0.7287	0.5793	0.6099	0.7155	0.7689	0.5846
	UOSTrack	0.6109	0.7067	0.5346	0.5618	0.6719	0.7377	0.5699
	MANTA	0.6417	0.7319	0.5843	0.6342	0.7564	0.7538	0.6594

Table 1. Performance comparison of tracking methods across four datasets. Best and second best results are highlighted in purple and blue respectively. For all the metrics, higher value indicates better performance.

centers, it ignores the accuracy of the predicted box dimensions. To jointly evaluate both factors, we introduce Center-Scale Consistency (CSC). For a predicted bounding box $b_p = (x_p, y_p, w_p, h_p)$ and the ground truth $b_g = (x_g, y_g, w_g, h_g)$, we first compute the normalized center error:

$$e_c = \frac{\|c_p - c_g\|_2}{\sqrt{w_g^2 + h_g^2}}, \quad c_p = \left(x_p + \frac{w_p}{2}, y_p + \frac{h_p}{2}\right), \quad (7)$$
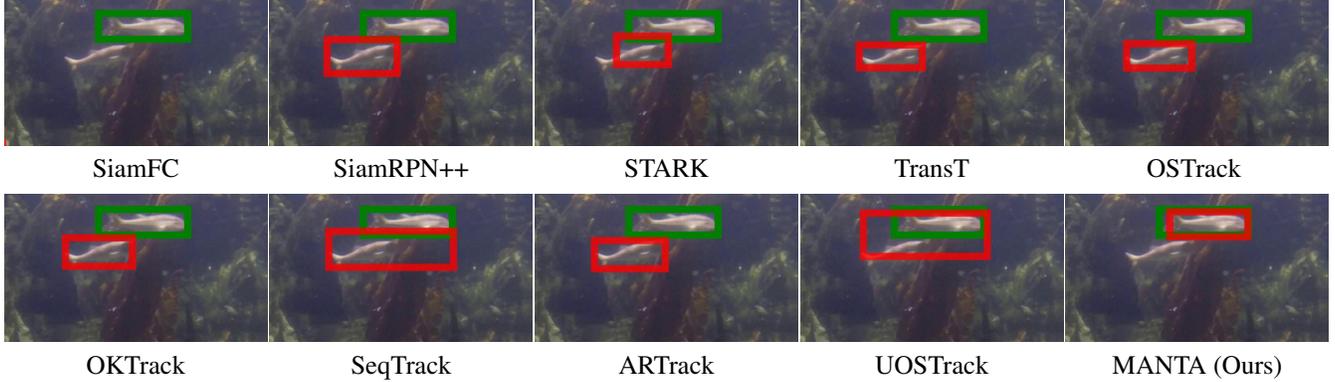
Figure 3. Comparison of tracking methods on frame 261 of the `WebUOT-1M_Test_000388` sequence from the WebUOT-1M [56] dataset. Tracker outputs are shown in red, and ground-truth boxes in green. Unlike other methods that either drift, miss the target, or track incorrect objects, MANTA consistently identifies and follows the correct object.

where $c_g = \left(x_g + \frac{w_g}{2}, y_g + \frac{h_g}{2} \right)$. We then define the relative scale errors as:

$$e_w = \frac{|w_p - w_g|}{w_g}, \quad e_h = \frac{|h_p - h_g|}{h_g}. \quad (8)$$

CSC is computed as the fraction of frames where both the center and scale errors fall below user-specified thresholds (τ_c, τ_s):

$$\text{CSC}(\tau_c, \tau_s) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left(e_c^i < \tau_c \wedge e_w^i < \tau_s \wedge e_h^i < \tau_s \right), \quad (9)$$

where $\mathbb{1}(\cdot)$ is the indicator function and the operator $\wedge$ denotes logical conjunction (AND). Intuitively, CSC measures how consistently the tracker predicts both the correct location and scale of the target.

Geometric Alignment Score (GAS): To provide a continuous measure of geometric fidelity, we propose the Geometric Alignment Score (GAS), which softly penalizes both misalignment and scale mismatch. Given the normalized center error e_c and scale error:

$$e_s = \frac{(w_p - w_g)^2 + (h_p - h_g)^2}{w_g^2 + h_g^2}, \quad (10)$$

we define:

$$\text{GAS} = \exp\left(-\frac{e_c^2}{\sigma_c^2} \right) \cdot \exp\left(-\frac{e_s}{\sigma_s^2} \right), \quad (11)$$

where σ_c and σ_s are tolerance parameters controlling sensitivity. GAS takes values in $(0, 1]$, with higher scores indicating better geometric alignment between prediction and ground truth.

In our experiments, we set $\tau_c = \tau_s = 0.2$ for CSC to require tight geometric alignment, and $\sigma_c = \sigma_s = 0.5$ for GAS to provide moderate tolerance for small misalignments.

4.4. Quantitative Results

Comparing Methods: We compare MANTA with nine state-of-the-art trackers, including Siamese networks (SiamFC [8], SiamRPN++ [30]), transformer-based models (STARK [52], TransT [14], SeqTrack [15], ARTrack [6], OSTrack [53]), and underwater-specific approaches (UOSTrack [33], OKTrack [56]).

Results across four benchmark datasets (Table 1) show MANTA consistently outperforms prior methods. On the UOT32 [28] dataset, it achieves 0.7306 Success AUC and 0.8829 Success@0.5, surpassing UOSTrack by 5–6%. Our metrics reveal even larger gains: +11.5% CSC@0.2 and +4.2% mGAS. On UTB180 [2], MANTA improves Success AUC (0.7328 vs. 0.7085 for ARTrack) and excels in alignment (0.8312 mGAS) and scale consistency (0.7746 CSC@0.2). For UW-COT220, it achieves the best mIoU (0.7631) and Success@0.5 (0.7036). On the large-scale WebUOT-1M, MANTA delivers consistent gains, including the top CSC@0.2 (0.6594). These improvements highlight the benefits of physics-informed contrastive learning and the vision-based re-identification, which enforces physical and track consistency, yielding good performance under low visibility, clutter, and scale variations.

4.5. Qualitative Results

Qualitative analysis demonstrates that MANTA maintains robust tracking across diverse underwater scenarios, including severe color distortions, low-visibility conditions, complex marine environments, and the presence of multiple distractor objects. The physics-informed representations enable consistent target identity preservation across frames, even under challenging conditions such as occlusion, low visibility, and appearance variations. As shown in Fig. 3, MANTA accurately identifies and follows the target fish, while competing methods such as STARK, TransT, OS-

Method	Succ. AUC	Prec. AUC	mIoU	mGAS	mCSC@0.2
None	0.7000	0.7280	0.7463	0.8560	0.6928
ResNet18	0.7046	0.7311	0.7653	0.8512	0.7221
Physics - T	0.7259	0.7576	0.7648	0.8767	0.7401
Physics - BL	0.7286	0.7596	0.7646	0.8802	0.7433
Physics - T + BL	**0.7306**	**0.7628**	**0.7653**	**0.8829**	**0.7438**

Table 2. Ablation study results on UOT32 dataset. We analyze the impact of encoder training in our MANTA framework. T and BL correspond to temporal and Beer-Lambert augmentations, respectively. Physics - T corresponds to the use of only temporal augmentaions.

Method	FPS	Parameters
SiamFC	30	2.3M
SiamRPN++	35	N/A
ARTrack	29	86M
OKTrack	52	92.1M
OSTrack	23	86M
SeqTrack	31	86M
STARK	17	87M
TransT	40	23M
UOSTrack	56	N/A
MANTA	37	45.4M

Table 3. Runtime and complexity analysis of tracking methods evaluated on the UOT32 [28] dataset (N/A - not available).

Track, and ARTrack drift towards incorrect objects, UOS-Track tracks multiple objects simultaneously, and SiamFC fails to detect the target altogether. In contrast, MANTA leverages an efficient re-identification mechanism to recover from occlusions and maintain the correct trajectory despite challenging underwater conditions. More detailed analyses and discussions are provided in the supplementary.

4.6. Ablation Studies

We validate the individual contributions of our proposed components through systematic ablations on the UOT32 dataset (Table 2). The baseline tracker achieves 0.7000 Success AUC and 0.7280 Precision AUC. Adding ImageNet-pretrained ResNet50 features offers only minor gains, underscoring the limited value of generic representations for UOT. Temporal contrastive learning improves performance to 0.7259 / 0.7576, while Beer–Lambert augmentations perform slightly better (0.7286 / 0.7596), showing the benefit of physics-informed independence. Combining both yields the best results: 0.7306 Success AUC, 0.7628 Precision AUC, and 0.7438 mCSC@0.2, with consistent improvements in mGAS. These results demonstrate that temporal and physics-based augmentations are complementary, and together enable the most robust underwater tracking.

4.7. Runtime and Complexity Analysis

Table 3 reports runtime (FPS) and parameter counts on UOT32 [28]. SiamFC is extremely lightweight (2.3M) but lacks capacity, while TransT offers a good trade-off (40

Figure 4. Runtime breakdown percentages of total runtime.

FPS, 23M). Heavy transformer trackers such as ARTrack, OSTrack, SeqTrack, and STARK (86–87M) run at 17–31 FPS, and OKTrack scales to 92.1M with 52 FPS, showing strong optimization but high memory cost. MANTA achieves 37 FPS with 45.4M parameters—about half the size of heavy transformer trackers while maintaining competitive speed. Compared to UOSTrack (56 FPS, complexity not reported), MANTA offers a clearer balance of efficiency and robustness. As shown in Fig. 4, most computation lies in detection (80.5%), while tracking (1.3%) and re-association (18.2%) are lightweight. Overall, MANTA provides an efficient speed–accuracy trade-off suited for real-time underwater tracking.

Limitations: Despite strong performance, MANTA has limitations. It relies on monocular depth estimation, which may be unreliable in texture-poor or repetitive scenes, though precise depth is unnecessary since we target invariance across attenuation patterns rather than precise physics modeling. The secondary association module uses fixed thresholds, limiting adaptability to varying conditions; adaptive thresholding could improve robustness. Finally, while effective for SOT, extending to MOT remains challenging due to the complexity of maintaining multiple identities under severe appearance distortions.

5. Conclusion

We introduced MANTA, a physics-informed framework for underwater object tracking that integrates Beer–Lambert–based augmentations into a dual-positive contrastive learning strategy. This design enables representations robust to temporal variations and underwater degradations. Integrated into a novel motion and vision-guided association tracker, MANTA yields consistent state-of-the-art performance across four underwater SOT benchmarks. In addition, our CSC and GAS metrics provide a more precise evaluation of geometric fidelity. Overall, MANTA demonstrates that embedding physical priors into self-supervised learning is a powerful direction for building robust and generalizable tracking systems in challenging real-world domains.

Acknowledgements

References

[1] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1682–1691, 2019. 2

[2] Basit Alawode, Yuhang Guo, Mehnaz Ummar, Naoufel Werghi, Jorge Dias, Ajmal Mian, and Sajid Javed. Utb180: A high-quality benchmark for underwater tracking. In *Proceedings of the Asian conference on computer vision*, pages 3326–3342, 2022. 1, 5, 6, 7

[3] Basit Alawode, Fayaz Ali Dharejo, Mehnaz Ummar, Yuhang Guo, Arif Mahmood, Naoufel Werghi, Fahad Shahbaz Khan, Jiri Matas, and Sajid Javed. Improving Underwater Visual Tracking With a Large Scale Dataset and Image Enhancement. 14, 2023. 1, 5

[4] Sara Aldhaheri, Giulia De Masi, Èric Pairet, and Paola Ardón. Underwater robot manipulation: Advances, challenges and prospective ventures. In *Oceans 2022-Chennai*, pages 1–7. IEEE, 2022. 2

[5] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. In *2012 IEEE conference on computer vision and pattern recognition*, pages 81–88. IEEE, 2012. 2

[6] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 7

[7] Chayan Banerjee, Kien Nguyen, Clinton Fookes, and Karniadakis George. Physics-informed computer vision: A review and perspectives. *ACM Computing Surveys*, 57(1):1–38, 2024. 2

[8] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 2, 7

[9] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 2, 4, 5

[10] Aditya Chandrasekar, Manogna Sreenivas, and Soma Biswas. Phish-net: Physics inspired system for high resolution underwater image enhancement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1506–1516, 2024. 2

[11] Haojie Chen, Zhuo Wang, Hongde Qin, and Xiaokai Mu. Self-supervised domain feature mining for underwater domain generalization object detection. *Expert Systems with Applications*, 265:126023, 2025. 1, 2

[12] Long Chen, Yuzhi Huang, Junyu Dong, Qi Xu, Sam Kwong, Huimin Lu, Huchuan Lu, and Chongyi Li. Underwater optical object detection in the era of artificial intelligence: Current, challenge, and future. *ACM Comput. Surv.*, 58(3), 2025. 1, 2

[13] Shujie Chen, Zhonglin Liu, Jianfeng Dong, Xun Wang, and Di Zhou. Representation alignment contrastive regularisation for multi-object tracking. *IET Computer Vision*, 19(1): e12331, 2025. 2

[14] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 2, 7

[15] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14572–14581, 2023. 2, 7

[16] Pierre-François De Plaen, Nicola Marinello, Marc Proesmans, Tinne Tuytelaars, and Luc Van Gool. Contrastive learning for multi-object tracking with transformers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6867–6877, 2024. 2

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[18] Mahmoud Elmezain, Lyes Saad Saoud, Atif Sultan, Mohamed Heshmat, Lakmal Seneviratne, and Irfan Hussain. Advancing underwater vision: a survey of deep learning models for underwater object recognition and tracking. *IEEE Access*, 2025. 1

[19] Girish Gaude and Samarth Borkar. Comprehensive survey on underwater object detection and tracking. *International Journal of Computer Sciences and Engineering*, 2018. 1

[20] Zhicheng Hao, Jun Qiu, Haimiao Zhang, Guangbo Ren, and Chang Liu. Umotma: Underwater multiple object tracking with memory aggregation. *Frontiers in Marine Science*, 9: 1071618, 2022. 1, 2, 5

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[22] Kaer Huang, Kanokphan Lertniphonphan, Feng Chen, Jian Li, and Zhepeng Wang. Multi-object tracking by self-supervised learning appearance model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3163–3169, 2023. 2, 3

[23] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 2

[24] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1769–1776. IEEE, 2020. 5

[25] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2, 2020. 1

[26] Muwei Jian, Nan Yang, Chen Tao, Huixiang Zhi, and Hanjiang Luo. Underwater object detection and datasets: a survey. *Intelligent Marine Technology and Systems*, 2(1):9, 2024. 2

[27] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. 2

[28] Landry Kezebou, Victor Oludare, Karen Panetta, and Sos S Agaian. Underwater object tracking benchmark and dataset. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–6. IEEE, 2019. 5, 6, 7, 8

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[30] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019. 2, 7

[31] Jiahua Li, Wentao Yang, Shishi Qiao, Zhaorui Gu, Bing Zheng, and Haiyong Zheng. Self-supervised marine organism detection from underwater images. *IEEE Journal of Oceanic Engineering*, 2024. 1, 2

[32] Xingkun Li, Yuhao Zhao, Hu Su, Yugang Wang, and Guodong Chen. Efficient underwater object detection based on feature enhancement and attention detection head. *Scientific Reports*, 15(1):5973, 2025. 2

[33] Yunfeng Li, Bo Wang, Ye Li, Zhuoyan Liu, Wei Huo, Yueming Li, and Jian Cao. Underwater object tracker: Uostrack for marine organism grasping of underwater vehicles. *Ocean Engineering*, 285:115449, 2023. 1, 7

[34] Yanli Li, Weidong Liu, Wenbo Zhang, and Le Li. Sequential state estimation based target tracking algorithm for unmanned underwater vehicle target grasping. *Engineering Applications of Artificial Intelligence*, 158:111313, 2025. 1

[35] Juntao Liang, Jun Hou, Weijun Zhang, and Yong Wang. Cftrack: Enhancing lightweight visual tracking through contrastive learning and feature matching. *arXiv preprint arXiv:2502.19705*, 2025. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[38] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International conference on image processing (ICIP)*, pages 3025–3029. IEEE, 2023. 2

[39] Simone Marini, Emanuela Fanelli, Valerio Sbragaglia, Ernesto Azzurro, Joaquin Del Rio Fernandez, and Jacopo Aguzzi. Tracking fish abundance by underwater image recognition. *Scientific reports*, 8(1):13748, 2018. 2

[40] Thomas G Mayerhöfer, Susanne Pahlow, and Jürgen Popp. The bouguer-beer-lambert law: Shining light on the obscure. *ChemPhysChem*, 21(18):2029–2046, 2020. 2, 3

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 5

[42] Karen Panetta, Landry Kezebou, Victor Oludare, and Sos Agaian. Comprehensive underwater object tracking benchmark dataset and underwater image enhancement with gan. *IEEE Journal of Oceanic Engineering*, 47(1):59–75, 2021. 1, 5

[43] Malte Pedersen, Daniel Lehotskỳ, Ivan Nikolov, and Thomas B Moeslund. Brackishmot: The brackish multi-object tracking dataset. In *Scandinavian Conference on Image Analysis*, pages 17–33. Springer, 2023. 1, 5

[44] Alan Preciado-Grijalva, Bilal Wehbe, Miguel Bande Firvida, and Matias Valdenegro-Toro. Self-supervised learning for sonar image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1499–1508, 2022. 1, 2

[45] Isaac Robinson, Peter Robicheaux, and Matvei Popov. Rf-detr. https://github.com/roboflow/rf-detr, 2025. SOTA Real-Time Object Detection Model. 4, 5

[46] Amirmohammad Shamaei, Jana Starcukova, and Zenon Starcuk Jr. Physics-informed deep learning approach to quantification of human brain metabolites from magnetic resonance spectroscopy data. *Computers in Biology and Medicine*, 158: 106837, 2023. 2

[47] Runmin Cong Suqi Li Wei Zhang Sam Kwong Shijie Lian, Hua Li. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1305–1315, 2023. 5

[48] Suhas Srinath, Aditya Chandrasekar, Hemang Jamadagni, Rajiv Soundararajan, and Prathosh AP. Undive: Generalized underwater video enhancement using generative priors. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9001–9012. IEEE, 2025. 2

[49] Donald F Swinehart. The beer-lambert law. *Journal of chemical education*, 39(7):333, 1962. 2, 3

[50] Yu Wei, Yi Wang, Baofeng Zhu, Chi Lin, Dan Wu, Xinwei Xue, and Ruili Wang. Underwater detection: A brief survey and a new multitask dataset. *International Journal of Network Dynamics and Intelligence*, pages 100025–100025, 2024. 1, 2

[51] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[52] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021. 7

[53] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022. 7

[54] Boxiao Yu, Jiayi Wu, and Md Jahidul Islam. Udepth: Fast monocular depth estimation for visually-guided underwater robots. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023. 3

[55] Di Yuan, Xiaojun Chang, Po-Yao Huang, Qiao Liu, and Zhenyu He. Self-supervised deep correlation tracking. *IEEE Transactions on Image Processing*, 30:976–985, 2020. 2, 3

[56] Chunhui Zhang, Li Liu, Guanjie Huang, Hao Wen, Xi Zhou, and Yanfeng Wang. Webuot-1m: Advancing deep underwater object tracking with a million-scale benchmark. *Advances in Neural Information Processing Systems*, 37:50152–50167, 2024. 1, 5, 6, 7

[57] Chunhui Zhang, Li Liu, Guanjie Huang, Zhipeng Zhang, Hao Wen, Xi Zhou, Shiming Ge, and Yanfeng Wang. Underwater camouflaged object tracking meets vision-language sam2. *arXiv preprint arXiv:2409.16902*, 2024. 1, 5, 6

[58] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. 2022. 2

[59] Yuan Zhou and Kangming Yan. Domain adaptive adversarial learning based on physics model feedback for underwater image enhancement. *arXiv preprint arXiv:2002.09315*, 2020. 2

[60] Xue-Feng Zhu, Tianyang Xu, Sara Atito, Muhammad Awais, Xiao-Jun Wu, Zhenhua Feng, and Josef Kittler. Self-supervised learning for rgb-d object tracking. *Pattern Recognition*, 155:110543, 2024. 2, 3

[61] Yaoming Zhuang, Jiaming Liu, Haoyang Zhao, Longyu Ma, Zirui Fang, Li Li, Chengdong Wu, Wei Cui, and Zhanlin Liu. A deep learning framework based on structured space model for detecting small objects in complex underwater environments. *Communications Engineering*, 4(1):24, 2025. 1